DOCUMENT RESUME

ED 365 148 FL 021 762

AUTHOR Porter, Don
TITLE Affective Factors in the Assessment of Oral Interaction: Gender and Status.
PUB DATE 91
NOTE 12p.; In: Sarinee, Anivan, Ed. Current Developments in Language Testing. Anthology Series 25. Paper presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (April 9-12, 1990); see FL 021 757.
PUB TYPE Reports - Research/Technical (143) -- Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communicative Competence (Languages); Interpersonal Relationship; Interrater Reliability; *Interviews; *Language Tests; *Oral Language; Predictor Variables; *Second Languages; *Sex Differences; Social Status; Speech Skills; *Student Attitudes; Student Characteristics; Testing

ABSTRACT

A discussion of oral language testing looks at the role of student attitudes, student and interviewer gender, and interviewer social status in the reliability of student assessments. Three small-scale studies investigating these factors are described. The first two involved only Arab students. In the first, it was found that students (all male) were generally given higher ratings by male interviewers. In the second, slightly larger study, both male and female students were rated higher by male interviewers. The third study's subjects were of varied cultural backgrounds, and an attempt was made to manipulate the students' perceptions of the relative status (high or neutral) of the interviewer. Preliminary results of this study suggest a similar, although less strong, tendency toward higher rating by male interviewers and, more surprising, toward higher ratings when the interviewer is not presented as high-status. Implications of such variables as possible predictors of performance are considered. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFFECTIVE FACTORS IN THE ASSESSMENT OF ORAL INTERACTION: GENDER AND STATUS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Weng

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Don Porter

INTRODUCTION

1 LINGUISTIC COMPLEXITY AND LANGUAGE TESTS

ED 365 148

In its internal structure and in its components, linguistic ability is extremely - if not infinitely - complex. Any attempt to summarize linguistic ability in the form of a description will necessarily have to consist of some form of simplification of the original complexity. Language tests are constructed on the basis of such simplifying descriptions of linguistic ability in general - what we might call linguistic 'models' - and are themselves devices for generating descriptions of the individual language user's ability in terms of the underlying model. So language tests, too, must simplify what they assess.

Sometimes the descriptions produced by a language test are in terms of numbers, eg '72%' or perhaps '59% in Writing, 72% in Reading' (although it is difficult to know what such descriptions of linguistic ability could mean, they are so abstract and relativistic); sometimes the descriptions are put in terms of verbal descriptions, eg:

> 'very little organisation of content; for the most part satisfactory cohesion; some inadequacies in vocabulary; almost no grammatical inaccuracies'
>
> (Based on criteria for Test of English for Educational Purposes: Weir, 1988)

But whatever form the description takes, the general headings under which the various aspects of the description fall are not God-given, inherent in the nature of language or linguistic ability,.so much as imposed on a continuum of confusing and unruly data by language specialists. Language ability does not fall neatly into natural pre-existing categories, but has to be forced into man-made categories with varying degrees of success. A description which aims for completeness by having special headings for all the bits which do not quite fit may well end up by being more complex than the original language ability being described - and the more complex the description gets, the less our brains are able to grasp it in its entirety: the less it means to us. A truly useful description

FL021762

BEST COPY AVAILABLE

of a language ability, then, will be one which leaves a great deal out! What such a description will do will be to focus on various features which are felt to be particularly salient and important. That is to say, it will be founded on a theoretical model - one which, in the features it chooses to highlight, and in the way it relates those features one to another, attempts to capture the essence of the language ability. The questions for a test, then, are: how elaborate a model should it be based on if it is to avoid the criticism that it leaves out of account crucial features of the language ability to be measured; and on the other hand how much complexity can it afford to report before it runs the risk of being unusable?

2 Communicative Language Testing

Testers vary in whether they claim to be producing or discussing communicative competence tests, communicative performance tests, or simply - and conveniently - communicative tests, and views of what those various terms imply also vary considerably. There is no widely accepted overall model of communicative proficiency used as a basis for this approach to language testing. Nevertheless, there is in Britain at least a fair degree of working consensus about the sorts of characteristics such tests ought to have. We may cite just the following few as being fairly typical:

(a) Tests will be based on the needs (or wants) of learners. It would be unreasonable to assess a learner's ability to do through English something which he has no need or wish to do. A principle such as this suggests that the different needs of different learners may call for different types of linguistic ability at different levels of performance; in principle tests incorporating this idea will vary appropriately for each new set of needs in the number and type of abilities they assess, and in their appraisal of what constitutes a satisfactory level of performance. Results will be reported separately for each ability in the form of a profile. We are thus immediately faced with a degree of test complexity at the points of test-content, assessment criteria, and report format.

(b) Tests will be based on language use in the contexts and for the purposes relevant to the learner. It is at least conceivable that any one of the linguistic ability types mentioned in the previous paragraph might be required in a number of distinct contexts crucial to the learner and for more than one distinct purpose in any given context. If varying context and purpose are seen as central features of natural communication, this

suggests that particular contexts and purposes require particular deployments of linguistic abilities. Both context and purpose will then need to be suitably incorporated in tests and will represent two further dimensions of complexity.

(c) Tests will employ authentic texts, or texts which embody fundamental features of authenticity. These 'fundamental features' may well include appropriate format and appropriate length, both of which will vary with the type of text. Concerning length in particular, longer texts are said to require types of processing different from those needed for shorter texts. Text authenticity then implies yet another dimension of complexity.

These characteristic features, together with others, reflect the assumption that, in Oller's (1979) terms, language ability is not unitary, but in fact very divisible.

Tests already exist which seek to embody all these and other features of natural communication for more or less well-defined groups of learners. The challenge is great and the difficulties formidable. Bachman (1990) has criticised such tests as suffering from inadequate sampling and consequent lack of generalizability of their results: the descriptions of ability yielded by the test, it is argued, refer only to the needs, contexts, purposes, text-types, etc. covered in the test; needs, contexts, purposes, etc. are so multifarious it is not possible to sample them adequately for all test-takers, and perhaps not even for a single test-taker.

In the light of the already-existing difficulties posed for test construction, and of such criticisms, and of the need for a useful, practical test to avoid excessive complexity, we must think very carefully indeed before proposing that tests should incorporate yet another level of complexity by including information on the effects of affective factors in the descriptions which they yield.

3 Affective Factors

Affective factors are emotions and attitudes which affect our behaviour. We may distinguish between two kinds: predictable and unpredictable.

Unpredictable: Most teachers will be familiar with the kinds of affective factor which produce unpredictable and unrepresentative results in language tests, eg. a residue of anger after a family row or a mood of irresponsibility after some unexpected good news on the day. The fact that such moods may weaken concentration, or may lead in some other way to learners not reflecting in their

performance the best that they are capable of, will obviously detract from the reliability of the description of abilities yielded by the test.

Clearly, if we can find ways of minimizing the effects of such unpredictable factors, we should do so. If the test is associated with a teaching programme, continuous assessment or a combination of continuous assessment with a formal test would be less likely to be affected by a single unrepresentative performance. On the other hand, if there is no associated teaching programme, and everything hangs on a single measure, we might try to eliminate from the subject matter of the test any topics which might be likely to touch on a raw nerve somewhere. For example, the Educational Testing Service carefully vets all essay topics for the Test of Written English for possible sources of unreliability, emotional associations being one such source.

Another possible route to eventual affect-free assessment might be to devise a programme of research to discover the kinds of test techniques which are least susceptible to emotional buffeting.

On the other hand, the attempt to eliminate emotional content from language tests, on whatever grounds, may be misconceived. Is it not the case that a fundamental and natural use of language is as a vehicle for messages with emotional associations? Imagine tests in which the learner is asked to react to or produce language with which he feels no personal involvement, and to which he feels no personal commitment. Would not such language be at least severely restricted in its range of content, and at most fundamentally unnatural? We are left in a dilemma: it is suggested that emotional content is a central feature of language use, but it is at the same time a potential source of unreliability.

The very unpredictability of such moods and emotions, however, means that there is a limit to the effectiveness of whatever measures we might take to deal with their effects. And if for some reason a learner does not feel like writing or talking, there is not a lot that we can do.

Predictable: There may be another set of affective factors which are predictable in their effects on the quality of communication, and which can therefore be built into a model of communicative performance. This is still an area of great ignorance and one worthy of much more research: we need to know what the predictable affective factors are, and what their sphere of influence is. It could be, for instance, that performance in spoken and written language is influenced by different sets of factors. But if we may from now on narrow our focus to performance in the spoken language, candidates for inclusion in the relevant set of predictable affective factors will include the age, status, personality-type (eg. 'out-going', 'reserved'), acquaintance-relationship, and gender of the participants. Let us now turn to three small studies which have aimed to shed some light on these questions, and to their implications.

4 Three Small Experimental Investigations

Investigation 1: Locke (1984) felt that the quality of spoken language elicited in an interview, or in any other face-to-face spoken interaction, might be crucially affected by features of the interlocutor - in the case of the interview, by features of the interviewer. Thus, if the interviewee was given interviewer 'a' he might do well, but if he was given interviewer 'b' he might do badly. Intuitively, her concern seemed reasonable, and was backed up by a wealth of anecdotal evidence. Yet most testing concern with unreliability in interview assessment focuses on lack of consistency in the assessor; attempts to strengthen reliability in the assessment of speaking ability focus on assessor training and the use of adequate and appropriate rating scales. Whilst the latter are undeniably important, the more fundamental point that the quality of spoken language performance may vary predictably with features of the interlocutor tends to go unnoticed. Research in this area is practically non-existent, although the results would be of importance beyond language testing for our understanding of the nature of linguistic performance.

Locke chose to consider the effect of the gender of the interviewer on the interviewee. Four male postgraduate Iraqi and Saudi students at the University of Reading were each interviewed twice, once by a male and once by a female interviewer. The four interviewers were all of comparable age. Two students were interviewed by a male interviewer first, and the other two by a female interviewer first; in this way it was hoped that any order effect could be discounted. Then, it was necessary for each interview to be similar enough to allow meaningful comparison of results, but not so similar that the second interview would be felt to be a simple repeat of the first, with a consequent practice effect. A 'same-but-different' format was therefore necessary. Each interview was given the same structure, and the general topic-area was also the same, but the specific content of the first and second interviews was different.

Each interview was video-recorded. Recordings were subsequently presented in a shuffled order, and assessed by one male and one female rater, each using two methods of assessment, one holistic (Carroll, 1980) and one analytic (Hawkey, 1982). In this way 16 comparisons of spoken language quality with male and female interviewers could be made.

Although the number of students was very small, the result was clear and provocative: there was an overwhelming tendency for students to be given higher ratings when interviewed by male interviewers. The tendency was evident in both scoring methods and there was a high level of agreement between the two raters.

Investigation 2: These results demanded both replication and deeper exploration. The writer therefore carried out a slightly larger investigation with

BEST COPY AVAILABLE

thirteen postgraduate Algerian students at Reading (11 males and two females). This time, interviewers were cross-categorized not only by gender, but also by whether or not the student was acquainted with them and by a rough categorization of their personality as 'more outgoing' or 'more reserved'. Once again, the age of interviewers was comparable.

As in Locke's study, order of presentation was controlled for, with six students being given the female, and seven the male interviewer first. Cutting across the male-female category, as far as was possible (given the odd number involved) roughly half of the students were acquainted with the interviewer in the first interview, and unacquainted in the second, with the other half of the students having the reverse experience; and again roughly half of the students received an 'outgoing' interviewer first, followed by a 'reserved' interviewer, with the remainder having the reverse experience. The interviews were again designed to be 'same-but-different', were video-recorded, shuffled, and rated using two methods of assessment.

The tendency observed in Locke's study, for students to be rated more highly when interviewed by men, was once again overwhelmingly found. The tendency was equally clear in both scoring methods, and the degree of difference was fairly constant at about .5 of one of Carroll's bands. Interestingly, neither of the other potential factors considered - acquaintanceship and personality-type - could be seen to have any consistent effect.

What was not clear from Locke's study and could only be trivially investigated in this one was whether any gender effect was the result of interviewees' reactions to males versus females, or to own-gender versus opposite-gender interviewers. In this respect, it was particularly unfortunate that more female students could not be incorporated in the study: female students of the same cultural background as the males were not available. Nevertheless, while expressing all the caution necessary when considering the results of only two students, the results for the two female students were interesting. For one of the women, no difference was observable by either scoring method with the male and female interviewers. The other woman was rated more highly when interviewed by the man. Neither woman could be seen to go against the trend established in the men.

A very tentative conclusion to be drawn from these two limited studies would seem to be that, in the interview situation at least, young adult male Arab students may have a consistent tendency to produce a higher quality of performance in spoken English when being interviewed by a man than when being interviewed by a woman.

If these studies really have, in a preliminary way, succeeded in detecting a predictable affective factor in spoken language performance, a number of further questions will need to be researched to clarify just what that affective factor is. As has been suggested above, it is still not clear whether what has been observed

concerns reaction to a male interviewer or to an own-gender interviewer. Further studies with female students would be needed in an attempt to answer this question.

Again, to what extent would this factor be restricted to Arab students? The emotive power of gender must surely pervade mankind, and thus such a gender-effect could be expected not only in any part of Europe but world-wide. On the other hand, Japanese colleagues say that they would not expect a gender-effect with Japanese students, but would not be surprised to find an age-effect, ie. we might expect students to achieve higher spoken-English ratings when interviewed by older interviewers, as such interviewers would be accorded greater respect. This interesting suggestion thus relates quality of performance in spoken language to the idea of degree of respect for the interviewer. A proposed gender-effect might thus be a manifestation of a more general 'respect' or 'status' effect. It might be that in many societies, but not all, men are accorded greater status than women, and that interviewees are moved to produce a higher quality of performance when confronted by high status in the interviewer. This suggests a need for a programme of research aimed at establishing and distinguishing between the effects of gender and status on quality of performance in spoken language.

Investigation 3: In an attempt to shed some light on this issue, a further small investigation was undertaken in Reading earlier this year. This is not yet complete, but preliminary indications are certainly of interest.

In this investigation, 16 postgraduate students were interviewed, coming from a variety of linguistic and cultural backgrounds. They included Arabs (Sudanese, Saudis, Yemenis and a Libyan), Japanese, Turks, and a Greek. Twelve students were male and four female.

As in the previous studies, each student was given two short 'same-but-different' interviews, one by a male interviewer, one by a female. Half of the students were interviewed by a male first, half by a female first, and all interviews were video-recorded.

The interviewers were roughly comparable in age, ranging from late twenties to early thirties. None of the interviewers was known to the students, and the personality of the interviewer was not controlled for. An attempt was made, however, to manipulate the status of each interviewer such that, in one interview the interviewer's status would be 'boosted' (high status), while in the

next it would not be (neutral status). Each interviewer (I) interviewed four students, thus:

	1st interview	2nd interview
Student # 1	Male I # 1 High status	Female I # 1 Neutral status
Student # 2	Female I # 1 High status	Male I # 1 Neutral status
Student # 3	Female I # 1 Neutral status	Male I # 1 High status
Student # 4	Male I # 1 Neutral status	Female I # 1 High status

The status of an interviewer was manipulated in the following way: if status was being 'boosted' the interviewer was introduced to the student by family name, and with academic titles where relevant (eg. Dr Smith). A brief description of the interviewer's affiliation and most important responsibilities was given. Most interviewers in this condition wore some formal items of clothing (eg. jackets for both men and women, ties for men, etc.) and the person introducing the interviewers maintained physical distance between himself and them. An attempt was made by the introducer to indicate deference through tone of voice. If status was not being boosted - the 'neutral status' condition - interviewers were introduced in a very friendly way, by first name only, as friends of the investigator and sometimes as graduate students in the Department of Linguistic Science. Jackets, ties, etc. were not worn, and in each introduction physical contact was made between the introducer and the interviewer, in the form of a friendly pat on the arm. Interviewers were instructed to 'be themselves' in both status conditions, their status being suggested to the student purely through the mode of introduction together with minor dress differences.

Videos of these interviews are currently being rated on holistic and analytic scales, as before. On this occasion, however, the holistic scales used are those developed by Weir for the oral component of the Test of English for Educational Purposes (see Weir, 1988), and in order to facilitate comparisons, the videos have not been shuffled. Multiple rating is being undertaken, with an equal number of male and female raters. Thus far, only two sets of ratings have been obtained, one by a male rater and one by a female.

While it is as yet much too early to draw any solid conclusions, some tentative observations are possible.

Firstly, the two raters agree closely, on both rating scales.

Secondly, there is a slight tendency on both rating scales and with both raters for students to achieve higher ratings when being interviewed by males, but this is by no means as clear-cut as in the earlier investigations, and on the analytic scales there is considerable disagreement between the raters on which criteria, or for which students, this tendency manifests itself. Nevertheless, some tendency is there.

Finally - and this, perhaps is the most surprising finding - there is some slight tendency on the analytic scale, and a more marked tendency on the holistic scale, for students to achieve higher ratings with interviewers who were not marked for high status!

If this latter suggestion is borne out when the analysis is complete, and if it is reinforced when more substantial studies are undertaken, it will raise some perplexing questions of interpretation. One possibility might be that it is not rather specific factors such as 'gender' or 'age', and not even a rather more general factor such as 'status' which affect the quality of language production directly, but some much more general, very abstract factor such as 'psychological distance'. Thus the more 'distant' an interlocutor is perceived to be, the poorer the ratings that will be achieved. All kinds of secondary factors might contribute to this notion of 'distance', in varying strengths, but an interlocutor who is 'same gender', 'same age', 'known to speaker', 'same status', etc. might be expected to elicit higher-rated language than one who is 'other gender', 'older', 'unknown to speaker'. 'higher status', etc.

Whatever the primary and secondary factors which ultimately emerge, if the nature and degree of effect can be shown to be consistent in any way for a specifiable group of speakers, this will suggest that a gender or status or psychological distance feature will have a good claim to be incorporated in models of spoken language performance for those speakers, and that tests of this performance will need to take such predictable factors into account.

Let us now consider what such 'taking account' might involve, and finally relate the whole issue to our underlying concern with the complexity of tests.

5 Taking Account of A Predictable Affective Factor

It is certainly not widespread current practice to take account of gender, status of participants, or 'distance' between them, in tests of oral interaction. The selection of interviewer or other type of interlocutor is normally a matter of chance as far as such factors are concerned, and no attempt is made to adjust results in the light of them. Some post-hoc adjustment of ratings would of

course be possible if the scale of an effect were known to be consistent. Thus a performance rating with a male interviewer could be converted to an equivalent rating with a female interviewer, or vice versa. But we would now be touching on very sensitive matters. This should not surprise us, and is not a unique byproduct of the particular affective factor chosen by way of illustration; the reader is reminded that affective factors are matters of emotion and attitude, and it is not only the testee who is subject to their effects!

The question arises, then, of whether it is appropriate to adjust ratings in such cases. What would be the standard to which particular results would be adjusted? Many people feel that the test should give the learner the chance to show the best that he can do, with the implication that the test results should report the learner's best achievement. But what if that were to mean for many groups of male learners that spoken language achievement with a female interviewer would be converted to a predictive description of what they would have been able to achieve if they had been interviewed by a man? Or something between the two? For many, this would not be an acceptable solution.

A slightly different approach would be to recognize that humanity incorporates gender differences, status differences etc., and that the quality of linguistic performance is conditioned by such factors. Care should therefore be taken to allow all major relevant factors to have full and appropriate play in each component of a language test, and the description of performance which would be the output of the test would be understood to be based on an incorporation of such factors. Thus it might be appropriate for all interviewees to be multiply interviewed, by interviewers of varying degrees and types of 'distance'.

This type of solution would have the added attraction of being able to deal with the effects of affective factors in cases where it was predictable that the factors would have a marked effect, but not predictable how great or in what direction the effect would be. Thus a 'distance' effect might be great in some individuals, or in people from some cultural backgrounds, but slight in others; great 'distance' might depress the quality of performance in some learners, but raise it in others.

It might at first glance appear that such a 'full play' solution would also have the attraction of making it unnecessary to do the research to find out what the significant factors would be. Simply replicate as closely as possible those situations in which the learner would be likely to find himself, and the appropriate affective factors would come into play of themselves. However the practicality of test construction and administration will inevitably require some simplification of reality as it appears in the test, some selection of the features to include - including what is felt to be important, excluding what is felt to be irrelevant. Research into what the significant affective factors are, the scale of their effects, and their field of operation (what topic-areas, what cultural backgrounds, etc) will be necessary to inform the selection process.

6 Affective Factors and The Complexity of Tests

We have considered in this paper only one small area of affectiveness. There are certain to be others which affect language performance, perhaps of much greater magnitude in their impact. The spoken language only has been considered; it may be that some or all of the factors affecting the spoken language will be shown to have significant effects on performance in the written language, too, to the same or different degrees. Alternatively, there may be a quite different set of affective factors for the written language. And in both media, the term 'performance' may be understood to involve both reception and production. The potential for test complexity if all are to be reflected in test content, structure and administration is quite awesome. Even the 'full-play' proposal of the previous section, related to a 'status' or 'distance' effect alone, would double at a stroke the number of interviewers required in any situation. Nevertheless, a description of a learner's linguistic performance which ignored this dimension of complexity would be leaving out of account something important.

But yes, in the end, practicality will have to win the day. Where the number of people taking the test is relatively small, and where the implications of the results are not critical in some sense, it is unlikely that affective factors will be, or could be, seriously and systematically taken into account. But where the test is a large one, where the results can affect the course of lives or entail the expenditure of large sums of money, and where specifiable affective factors are known to have significant effects on linguistic performance, it would be dangerous to ignore them.

REFERENCES

Bachman, L. 1990. Fundamental Considerations in Language Testing. London: CUP.

Carroll, B J. 1980. Testing Communicative Performance: An Interim Study. Oxford: Pergamon.

Hawkey, R. 1982. Unpublished Ph D. Thesis, University of London.

Locke, C. 1984. Unpublished MA Project, University of Reading.

Oller, J W 1979. Language Tests at School. London: Longman.

Weir, C. 1988. Communicative Language Testing. University of Exeter Press.